Preserving Geospatial Data: The National Geospatial Digital Archive's Approach

Greg Janée; University of California at Santa Barbara; Santa Barbara, CA, USA

Abstract

The National Geospatial Digital Archive (NGDA) is one of eight initial projects funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP). The project's overarching goal is to answer the question: How can we preserve geospatial data on a national scale and make it available to future generations? This paper summarizes the project's work in four areas: analysis of the characteristics of geospatial data relevant to preservation; elucidation of the "relay" principles of long-term preservation; development of an OAIS-compliant archive system; and development of a wiki- and repository-based format registry.

Introduction

The National Geospatial Digital Archive (NGDA),[1] a partnership between the Map & Imagery Laboratory, Davidson Library, at the University of California at Santa Barbara, and Branner Earth Sciences Library at Stanford University, is one of eight initial projects funded by the Library of Congress's National Digital Information Infrastructure and Preservation Program (NDIIPP).[2] The project's overarching goal is to answer the question: How can we preserve geospatial data on a national scale and make it available to future generations? Work on the project began in earnest in 2005 and immediately led to several new questions being posed:

- What are the characteristics of geospatial data that impact preservation?
- Given a desire to preserve information for a century or longer—a period of time far exceeding the lifetimes of the applications, platforms, and people involved in the information's creation—is there any preservation architecture, or are there at least any general design principles or best practices, that can carry the information through a century of unforeseeable technological and social change?
- Given a desire to preserve information on a large scale, can we define a minimal level or minimum standard of preservation that has a high chance of being achieved over the course of a century, without interruption or discontinuity, so that the information remains (at least potentially) as useful as when it was created, despite unforeseeable fluctuations in available resources devotable to the information's curation over time, and fluctuations in interest in the information and in the information's perceived value?

This paper summarizes NGDA's work in answering these questions. In the next section we list characteristics of geospatial data relevant to preservation. In the subsequent two sections, we elucidate three principles of long-term preservation and describe a prototype archive system built by NGDA that satisfies those principles. Finally, we describe NGDA's work in developing a wiki- and repository-based format registry.

Geospatial data characteristics

Geospatial data refers to the wide variety of scientific and government-produced datasets that have a geographic component, and that can typically be viewed as representing a portion of the Earth's surface in some way. This class of information encompasses remote-sensing imagery, aerial photography, maps, data produced by both fixed and mobile geographically-embedded sensors, and data created and processed by GIS (Geographic Information System) tools.

The following are some characteristics of geospatial data that are relevant to its preservation.

No uniform data model. Geospatial data spans a wide variety of data organizations: vector and raster; topological and non-topological; over domains both discrete and continuous. Geospatial applications and file formats support differing subsets and aspects of these data organizations, and to varying degrees. One attempt at defining a universal, public data model for geospatial data has been made, the USGS SDTS format,[3] but it has failed to achieve widespread adoption. As a consequence, it is not possible to speak of "geospatial data" as a single type of quantity that can be handled by multiple, functionally equivalent applications and formats.

Proprietary formats. Many geospatial formats, particularly GIS formats, are proprietary and therefore closely tied to applications. Furthermore, as is typical with formats driven by marketplace competition, they are frequently subject to backwardly incompatible revisions over time.

Multiple granule sizes. In contrast to textual information, which has been successfully modeled using multi-page, (hyper)textual documents as the sole granule size, geospatial data is regularly processed at varying granule sizes. The granule sizes range from individual features having geographic location, geometry, and related attributes; to homogeneous, thematic layers of features; to integrated, heterogeneous databases. Data can be aggregated, disaggregated, and operated on with some fluidity. Each of these granularities has its uses, affords different functionality, and poses different preservation challenges. As a consequence, there is no single preservation problem for geospatial data; instead, choosing which level or granule size to address, and

---

[1] http://www.ngda.org/
[2] http://www.digitalpreservation.gov/

[3] http://mcmcweb.er.usgs.gov/sdts/

therefore identifying the preservation problem(s), is a first step of the process.

Relational data systems. Geospatial data managed by GIS tools is more and more often being stored in "geodatabases": relational databases with geographic extensions. The virtue of the geodatabase—that it provides a unified, seamless environment in which to store complex relationships among heterogeneous features—is also a bane for preservation, as it means that it is often not possible to extract individual components out of the database without losing information. And geodatabases inherit all the problems of preserving relational databases: the need to take snapshots of running database systems; storage of snapshots in proprietary database dump formats; complex dump formats; and large, monolithic snapshot files.

Large size. The size of geospatial data is large by any measure, with datasets commonly having gigabyte granularities and with some datasets growing by terabytes per day.

Long-lived programs. Geospatial datasets can be long-lived: satellite-based sensor programs may run for years, even decades. As a consequence, it becomes necessary to begin archiving datasets long before they are "finished." Traditionally this has been addressed by binding datasets to storage systems that inevitably become obsolete even within the program's lifetime, but archival systems of the future that hope to lower both the cost of preservation and the risk of information loss will need to be designed to allow easy turnover and handoff of ever-evolving components and technologies.

Extensive context. Capturing and preserving enough of the context surrounding geospatial data to support the data's future interpretation and use can be challenging. Whereas format information by itself is sufficient to support future renderability of multimedia documents (e.g., knowledge of the PDF format is sufficient to render PDF documents, and therefore usability by humans), geospatial data can require much more, and more complex, contextual information. Using remote-sensing imagery in scientific modeling requires detailed knowledge of platform and sensor characteristics, and in many cases calibration and processing steps as well. Strictly speaking, such contextual information constitutes metadata, but in practice, being voluminous, it is not handled as such (for example, it is not stored in metadata records bundled with the data).

Implicit context. In many cases, the context surrounding geospatial data is implicit and embedded in small, relatively insular scientific communities.

Dynamic data. Some datasets, particularly Climate Data Records (CDRs), may need to be periodically reprocessed from source datasets in response to corrections and improvements in calibration and Earth models. Thus the context for these datasets must include not only information for their use, but information for their (re)processing as well, including software, algorithms, workflows, ancillary calibration tables, and other artifacts. And, in addition to simply storing such information, it must be possible to re-execute workflows, implying that lineage relationships between datasets and source datasets must be actively maintained. In the larger view, science datasets reside in a dynamic ecosystem of related datasets, and to preserve a dataset means to preserve the dataset's ability to function in that ecosystem.

From these characteristics we conclude that several challenges arise in preserving geospatial data over those already imposed by the general digital preservation problem. Whereas a multimedia document typically resides within a single file, geospatial data may reside in complex, multi-file objects. Whereas the interpretation of a PDF document may be defined by the format label "PDF," and in turn by an entry in a central format registry, geospatial data may require extensive, product-specific context to interpret. Whereas a thesis or journal article is fixed upon publication, geospatial data can remain dynamic indefinitely due to the lifetime of the generating program and the need to be periodically reprocessed.

Relay-supporting preservation architectures

We now turn to NGDA's work on preservation architectures. In thinking about how information can be preserved, it is natural to focus on the *system* that will house the information: a system must be built to hold the information and make it accessible; the system's purpose is (at least in part, if not wholly) to preserve the information; and hence, it is tempting to think, by building the system, the preservation of the information will have been addressed. This line of thinking is particularly attractive if the system supports preservation-related functionality such as format migration.

But if our goal is to preserve the information for a century or longer, it is evident that any system, no matter how well-designed or well-supported or preservation-supporting, is destined to become obsolete and unsupportable long before the century mark. Currently, storage systems become obsolete within a few years; storage media technologies, within a decade. At the next level up, in NGDA's experience in running libraries and data centers, we have found it very difficult to keep any type of data management system (repository system, digital library, etc.) running for even a decade. And at the highest level, curators and institutions themselves come and go over time. Few institutions can guarantee their own existence over a century, let alone their ability to continuously preserve and curate any particular piece of information. Instead, as Chris Rusbridge of the Digital Curation Centre has observed, long-term preservation is more likely to resemble a series of short-term guarantees measured in decades or less.

Thus we argue that preservation takes the form of an extended "relay" over time [5]. Preserving digital information for a century will require a series of handoffs, occurring repeatedly at many levels: between different types of media and storage subsystems, different object frameworks and organizational schemes, different repository systems, different institutions and policy regimes, and different, diverse application communities. The design of such an archive relay for digital information must focus on achieving the kind of interoperability that maximizes the ease with which such handoffs can successfully be made, in spite of the heterogeneity that will be introduced at many steps along the way [3].

Furthermore, the problems in making successful handoffs are likely to be exacerbated over time as archives of the future find themselves curating older and older digital information. Given our short digital history, most archives today are in the fortunate position of working with recently-created information; that is, with information types that are still current and well-understood in their respective communities. But if we consider our 100-year reference timespan, archives from the middle to the end of that

span will be faced with curating information for which all links to the original creators and context have been severed. To see this, one only has to consider the challenges, in the year 2009, of curating digital materials created in 1959, or 1909.

Architectural principles

NGDA has identified three architectural design principles that extend the recommendations made by the OAIS standard [2] and that we believe are necessary to support preservation of information across long-lived chains of curators and preservation systems.

Relay principle

The "relay" principle states: *a preservation system should support handoff of its archived content to the next preservation system in succession; that is, the preservation system should support its own migration.* (Note that we're distinguishing migration of the *system* itself here from migration of archived content *within* the system, e.g., file format migration.) Furthermore, the system should support its own migration at the archive, repository system, and storage system levels independently, to accommodate the different rates at which handoffs occur at these different levels and the different challenges that arise in each case.

If we take as a running, simplified example a user managing a set of photographs on a personal computer, with the photo management program (iPhoto, Picasa, etc.) playing the role of the repository system, then this principle states that the management program should support migration of the user's photo library to another, different photo management program. (This principle is perhaps analogous to recent calls for Web 2.0 data ownership and portability principles as exemplified by the DataPortability Project[4].) This principle also requires that, independently, the photo management program support handoff of just the storage of the photo library, for example, from one disk or computer or storage system to another.

Fallback principle

The "fallback" principle states: *a preservation system should support some form of handoff of its content even in the situation when the system itself is no longer functional.*

The OAIS standard describes data movement in terms of submission and dissemination information packages (SIPs and DIPs) and ingest and export functions. The transfer of information packages is certainly one means of migrating content across systems, but we believe a preservation risk is introduced if it is the *only* means. At the time of the handoff, the old preservation system may no longer be running; it may need to be reinstalled, but it may no longer be supported; the platform operating system may have been compromised and need to be reinstalled, but only a newer version is available which is incompatible with the version of the preservation system; the curating institution may want or need to move to a new preservation system, but not have the resources to keep the old preservation system running; etc., etc. There are many such scenarios, and we can summarize them all by

observing that a handoff may be required precisely *because* the old system can no longer be maintained or supported. In such cases, the preservation system should provide a "fallback" means of migration, e.g., by storing all archived content and preservation-related metadata in open (non-proprietary) files stored in a simple directory structure in a filesystem. In our photo example, iPhoto stores photos as native JPEG files in a readily understandable directory hierarchy, and while it stores metadata in various proprietary files, it also keeps all metadata[5] synchronized and exported in a human-readable XML file. Thus iPhoto satisfies the fallback principle. One can remove iPhoto entirely and still be left with a set of files that can be ingested into a new photo management program with relatively little and easy scripting. There's still a dependency on the viability and interpretability of the filesystem, of course. But there are many situations when a filesystem (that is, the files within the filesystem) can be recovered from an otherwise failed system, and too, filesystems have proved to be a remarkably resilient and roughly interoperable feature of computer systems.

Resurrection principle

The "resurrection" principle states: *a preservation system should allow archived information to lapse out of usability as a cost-saving measure, and should store and preserve sufficient metadata and contextual information to support future resurrection of full access and use of the information.*

There are many preservation risks to be addressed, from storage loss to format obsolescence, but a key risk that may be overlooked is the possibility of a lack of resources (time, money, personnel, expertise) to properly curate information. For any given piece of information, we cannot assume that the information will be maintained in a baseline usable state, let alone fully curated, at every point of its existence. The perceived value of information changes over time; archive resources inevitably change over time; and there may be periods of time during which the upkeep of the information cannot be supported or justified. Furthermore, the risk of insufficient resources is acutely significant at handoff points as described previously under the fallback principle, particularly handoffs between archive systems and between institutions. This risk can be mitigated by allowing the information to drop into a low-cost, unusable state (i.e., a state incurring the cost of bit storage only) with the proviso that sufficient contextual information is included and preserved to allow programmers and domain specialists of the future to resurrect access and usability as desire and resources permit.

In our photo example, this principle requires that the photo management program store, at minimum, format specifications for the JPEG and XML formats and any other formats required to reconstruct the photo library and viewability of the photos. In practice, of course, no photo management program does this. For widely-used formats such as JPEG and XML, it may be acceptable to refer to entries in central format registries such as GDFR[6] or

---

[4] http://dataportability.org/

[5] In iPhoto versions 5 and 6, *almost* all metadata; collection descriptions are not exported to the XML file.

[6] http://www.gdfr.info/

PRONOM[7] or the Library of Congress's "Sustainability of Digital Formats" website[8]. However, we note that none of these registries (as of this writing) stores format specifications or has the facilities or system architecture to curate format specification documents.

Combined with the fallback principle, the resurrection principle implies that format and other contextual information must be retrievable from a preservation system that is no longer functioning, e.g., by being stored as open files in a filesystem.

The NGDA archive system

NGDA has implemented a testbed archive system that satisfies the above principles.

At the storage level NGDA has experimented using Logistical Networking (LN) [1], a technology that provides seamless migration and replication of data across storage systems. LN is the most explicit attempt to date to apply the Internet's architectural approach to storage. The key to the Internet's design is an "hourglass" architecture, at the narrow waist of which is a highly generic, common service—the IP protocol for best-effort datagram delivery—that mediates between basic shared physical resources (network bandwidth in the Internet's case) and the applications that want to use those resources. Protocols built on top of IP, such as TCP, provide higher-level functionality such as reliable communication and persistent connections. LN's basic elements closely track this design. At the narrow waist of LN is the Internet Backplane Protocol (IBP), which mediates (only) best-effort, relatively short-term storage leases. Higher-level protocols built on IBP provide persistent and replicated storage, abstractions such as files and filesystems, and so forth.

Logistical Networking, if adopted as widely as other Internet protocols, could change how we conceive of and use storage. It would take functionality that is currently available on local scales and often delimited by proprietary boundaries—bit movement and replication automated to the extent that storage actions are expressible as simple, declarative policy and ownership changes—to a global scale.

At the repository level, NGDA has developed an archive system that implements a logical data model capable of modeling the kinds of object structures and relationships we have observed in geospatial data. Details of the data model are given in [4], and we note here only that the data model is broadly similar to OAI-ORE [7], though more constrained and focused on the requirements of long-term preservation. For example, while OAI-ORE mandates no specific relationships between Web resources, the NGDA data model mandates that each archival object (and each component thereof) have at least one relationship that defines the semantics of the object or component; furthermore, in a kind of recursion, the target of the relationship must be another archival object.

NGDA has also defined a physical data model, corresponding to the logical data model, that specifies how archival objects and components are laid out in a filesystem, how XML manifests for archival objects are placed and named in the filesystem, and how archival object identifiers are bidirectionally mapped to filesystem pathnames.

The logical data model, being able to capture the necessary context, satisfies the resurrection principle, while the physical data model satisfies the fallback principle. Thus, as with the iPhoto example, the NGDA archive system can be entirely removed and one is still left with an orderly set of files to support system-level migration. We are currently examining how, and with what modifications or extensions, other repository systems, namely DSpace[9] and Fedora,[10] can support the preservation principles we have outlined.

At the archive level, NGDA's approach agrees with the AIHT experiment's advocacy of a data-centric approach to migration: "A data-centric strategy assumes that the interaction between institutions will mainly be in the passing of a bundle of data from one place to another—that data will leave its original context and be interpreted in the new context of the receiving institution." [6] Every archive will have associated policies it imposes and services it provides, but it is primarily (and perhaps only) the data that will be propagated in the relay across time. Thus we have started investigating the notion of a "whole-archive descriptor" that describes (only) an archive's root crawl point(s) and any dependencies on other archives, registries, or identifier resolution services needed to support interpretation of the archive's content.

Wiki- and repository-based format registry

Preserving any type of digital information requires, in turn, preserving the information's context to support future interpretability. For many types of information, knowledge of the information's file format is sufficient, but as mentioned previously, geospatial data can require context beyond file formats. The data's context may be voluminous, as it is in the case of remote-sensing imagery; or it may be complex, dynamic, and evolving to support data reprocessing, as it is in the case of Climate Data Records.

NGDA initially looked to the digital library community's efforts in the area of context preservation, which were focused on the development of format registries, the most well-known being the GDFR, still under development, and PRONOM, now operational but still sparsely populated. The Library of Congress's "Sustainability of Digital Formats" website is effectively a format registry as well. These efforts all share several characteristics:

- They have all been targeted at relatively ubiquitous document and multimedia formats, as opposed to more narrowly-scoped formats used within specific communities.
- They have all adopted what might be called a "portal" view of registries, which is to say that the registry does not itself serve as a repository for format specifications, but instead simply refers to specifications and format-related websites on the Web. Facilities for storing simple files in these registries are primitive at best; facilities for storing and curating complex compound objects, software, etc., are nonexistent.
- They are relatively closed environments.

[7] http://www.nationalarchives.gov.uk/pronom/
[8] http://www.digitalpreservation.gov/formats/

[9] http://www.dspace.org/
[10] http://www.fedora-commons.org/

In summary, current registry efforts leave the issue of preservation of contextual information itself largely unaddressed.

NGDA has addressed this issue in two ways. First, NGDA's logical data model for archival objects seamlessly integrates data repositories and format registries. In the NGDA architecture, formats (and other types of semantics-defining objects) are themselves represented as archival objects, and thus may be curated as such. Thus the NGDA data model replaces the bifurcated view of archives of objects on the one hand referencing registries of formats on the other, with a unified view of inter-related archival objects residing and referencing each other across a federation of archives. In NGDA's view, a format registry is also an archive, just one that stores format metadata.

Second, NGDA has developed and experimented with a prototype wiki environment that allows communities to collaboratively author and gather format and other contextual information. Archive curators, using NGDA-supplied software, correlate and mediate the wiki view of this contextual information with an archive view based on representation of formats and other contextual information as archival objects as described above. In this way the system balances the needs of communities (flexibility, openness) and the needs of curators (structure, uniformity, control). The wiki interface is being exercised now within the project to populate the registry with geospatial format information.

Acknowledgements

References

[1] Micah Beck, Terry Moore, & James S. Plank (2002). An End-to-end Approach to Globally Scalable Network Storage. *ACM SIGCOMM Computer Communication Review* **32**(4) (October 2002):339–346. doi:10.1145/964725.633058

[2] Consultative Committee for Space Data Systems (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1, Blue Book (January 2002). http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] Margaret Hedstrom (2001). Exploring the Concept of Temporal Interoperability as a Framework for Digital Preservation. *Third DELOS Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries* (Darmstadt, Germany; September 8–9, 2001). http://www.ercim.org/publication/ws-proceedings/DelNoe03/10.pdf

[4] Greg Janée, Justin Mathena, & James Frew (2008). A Data Model and Architecture for Long-term Preservation. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (Pittsburgh, PA; June 16–20, 2008):134–144. doi:10.1145/1378889.1378912

[5] Greg Janée, James Frew, and Terry Moore (2008). Relay-supporting Archives: Requirements and Progress. *Proceedings of the 4th International Digital Curation Conference* (to appear). http://www.alexandria.ucsb.edu/~gjanee/archive/2008/idcc.pdf

[6] Clay Shirky (2005). *Library of Congress Archive Ingest and Handling Test (AIHT) Final Report*. http://www.digitalpreservation.gov/library/pdf/ndiipp_aiht_final_report.pdf

[7] Herbert Van de Sompel & Carl Lagoze (2007). Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication. *CTWatch Quarterly* **3**(3) (August 2007):32–41. http://www.ctwatch.org/quarterly/articles/2007/08/interoperability-for-the-discovery-use-and-re-use-of-units-of-scholarly-communication/

Author Biography

Greg Janée is a digital library research specialist at both the Institute for Computational Earth System Science (ICESS) and the Map & Imagery Laboratory, Davidson Library, at UC Santa Barbara. He is currently technical leader of the NGDA project. Prior to NGDA, he was technical leader of the Alexandria Digital Library (ADL) and Alexandria Digital Earth Prototype projects; principal developer of the ADL digital library software; and principal author of the ADL gazetteer and thesaurus protocols. He holds a BS in mathematics and an MS in computer science, both from UC Santa Barbara.